Database Design and Implementation

CS 645

Data provenance

Provenance

provenance, n.

The fact of coming from some particular source or quarter; origin, derivation [Oxford English Dictionary]

- Data provenance / lineage
  - [BunemanKhannaTan'01]: aims to explain how a particular result was derived.
- Data-intensive science
  - Worry about provenance

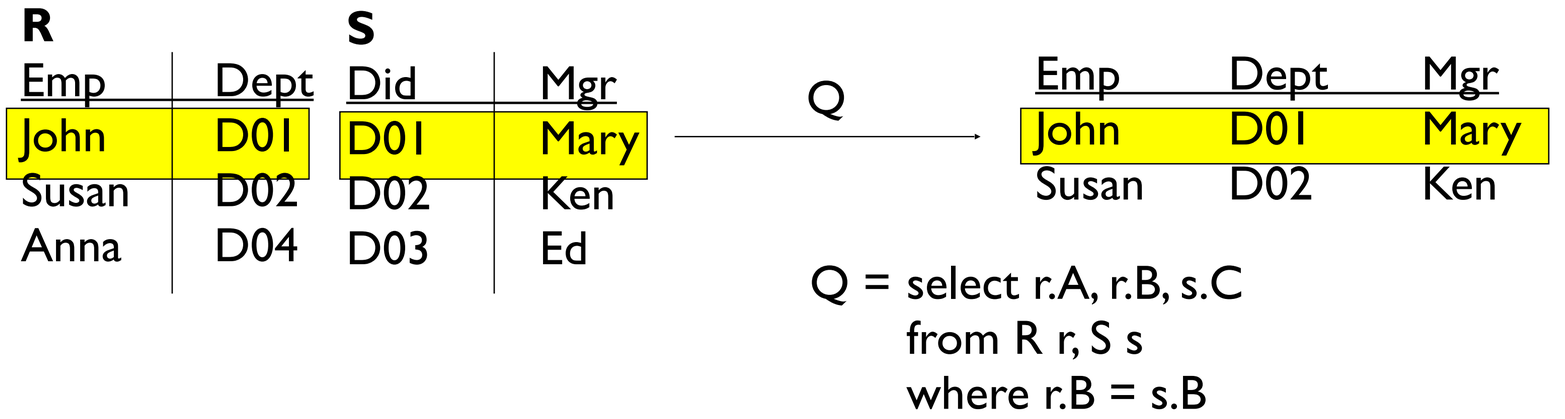
Motivation

- Data integration [WangMadnick90, LeeBressanMadnick98]
- Data Warehousing [CuiWidonWiener00]
- Scientific Data Management [BunemanKhannaTan01]

  - Determines trust on results
  - Ensure reliability, quality of data
  - Repeatability/verifiability
  - Avoid effort duplication
  - Understanding transport of annotations

Example of data provenance

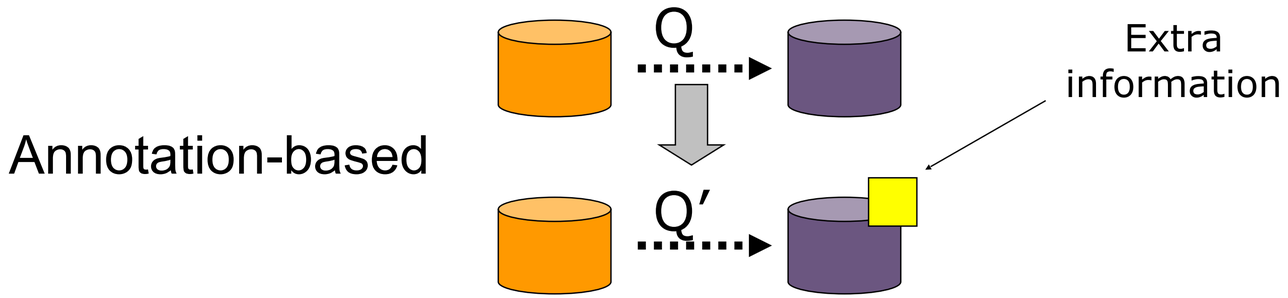
- A typical question:
  - For a given database query Q, a database D and a tuple t in the output of Q(D), which parts of D "contribute" to t?



  - The question can be applied to attribute values, tables, etc.

Two approaches

- Eager or annotation-based
  - Changes the transformation from Q to Q’ to carry extra information
  - Source data not needed after transformation



- Lazy or non-annotation based
  - Q is unchanged
  - Good when extra storage is an issue
  - Recomputation and access to source required

Types of provenance

- Why
  - *“What DB tuples contribute to the presence of each result tuple?”*
- How
  - *“By what process is each output tuple produced from the DB instance?”*
- Where
  - *“Where (from what attribute of what tuple) does each output tuple value come from?”*

Why-provenance example

Agencies

	name	based_in	phone
t_1:	BayTours	San Francisco	415-1200
t_2:	HarborCruz	Santa Cruz	831-3000

ExternalTours

	name	destination	type	price
t_3:	BayTours	San Francisco	cable car	$50
t_4:	BayTours	Santa Cruz	bus	$100
t_5:	BayTours	Santa Cruz	boat	$250
t_6:	BayTours	Monterey	boat	$400
t_7:	HarborCruz	Monterey	boat	$200
t_8:	HarborCruz	Carmel	train	$90

Q:

```
SELECT DISTINCT a.name, a.phone
FROM   Agencies a, ExternalTours e
WHERE  a.name = e.name
       AND e.type='boat'
```

Result of Q_1:

name	phone
BayTours	415-1200
HarborCruz	831-3000

- *Lineage* for an output tuple *t* is a subset of the input tuples which are *relevant* to the output tuple

Agencies

	name	based_in	phone
t_1:	BayTours	San Francisco	415-1200
t_2:	HarborCruz	Santa Cruz	831-3000

ExternalTours

	name	destination	type	price
t_3:	BayTours	San Francisco	cable car	$50
t_4:	BayTours	Santa Cruz	bus	$100
t_5:	BayTours	Santa Cruz	boat	$250
t_6:	BayTours	Monterey	boat	$400
t_7:	HarborCruz	Monterey	boat	$200
t_8:	HarborCruz	Carmel	train	$90

Q:

```
SELECT DISTINCT a.name, a.phone
FROM    Agencies a, ExternalTours e
WHERE  a.name = e.name
           AND e.type='boat'
```

Result of Q_1:

name	phone
BayTours	415-1200
HarborCruz	831-3000

Lineage: {t_1, t_5, t_6}

Problem: Not very precise.
e.g., lineage above does not specify that t_5 and t_6 do not both need to exist.

Why provenance

Agencies

	name	based_in	phone
t_1:	BayTours	San Francisco	415-1200
t_2:	HarborCruz	Santa Cruz	831-3000

ExternalTours

	name	destination	type	price
t_3:	BayTours	San Francisco	cable car	$50
t_4:	BayTours	Santa Cruz	bus	$100
t_5:	BayTours	Santa Cruz	boat	$250
t_6:	BayTours	Monterey	boat	$400
t_7:	HarborCruz	Monterey	boat	$200
t_8:	HarborCruz	Carmel	train	$90

Q:

```
SELECT DISTINCT a.name, a.phone
FROM    Agencies a, ExternalTours e
WHERE  a.name = e.name
           AND e.type='boat'
```

Result of Q_1:

name	phone
BayTours	415-1200
HarborCruz	831-3000

Lineage: {t_1, t_5, t_6}

Witness of *t*: Any subset of the database sufficient to reconstruct tuple *t* in the query result.

$\{t_1, t_5\}$ $\{t_1, t_6\}$ $\{t_1, t_2, t_6, t_8\}$

Witness basis: Leaves of the "proof tree" showing how result tuple *t* is generated

$\{\{t_1, t_5\}, \{t_1, t_6\}\}$

The view deletion problem

- *D* a database instance and *V=Q(D)* a view defined over *D*.
  - *Find a set of tuples ΔD to remove from D so that a specific tuple t is removed from the view*
  - *Minimize the number of side-effects in the view*
    - *View side-effect problem*
      - *Hard: queries with joins and projection or union*
      - *PTIME: the rest*
  - *Minimize the number of tuples deleted from D*
    - *Source side-effect problem*
      - *Same dichotomy*

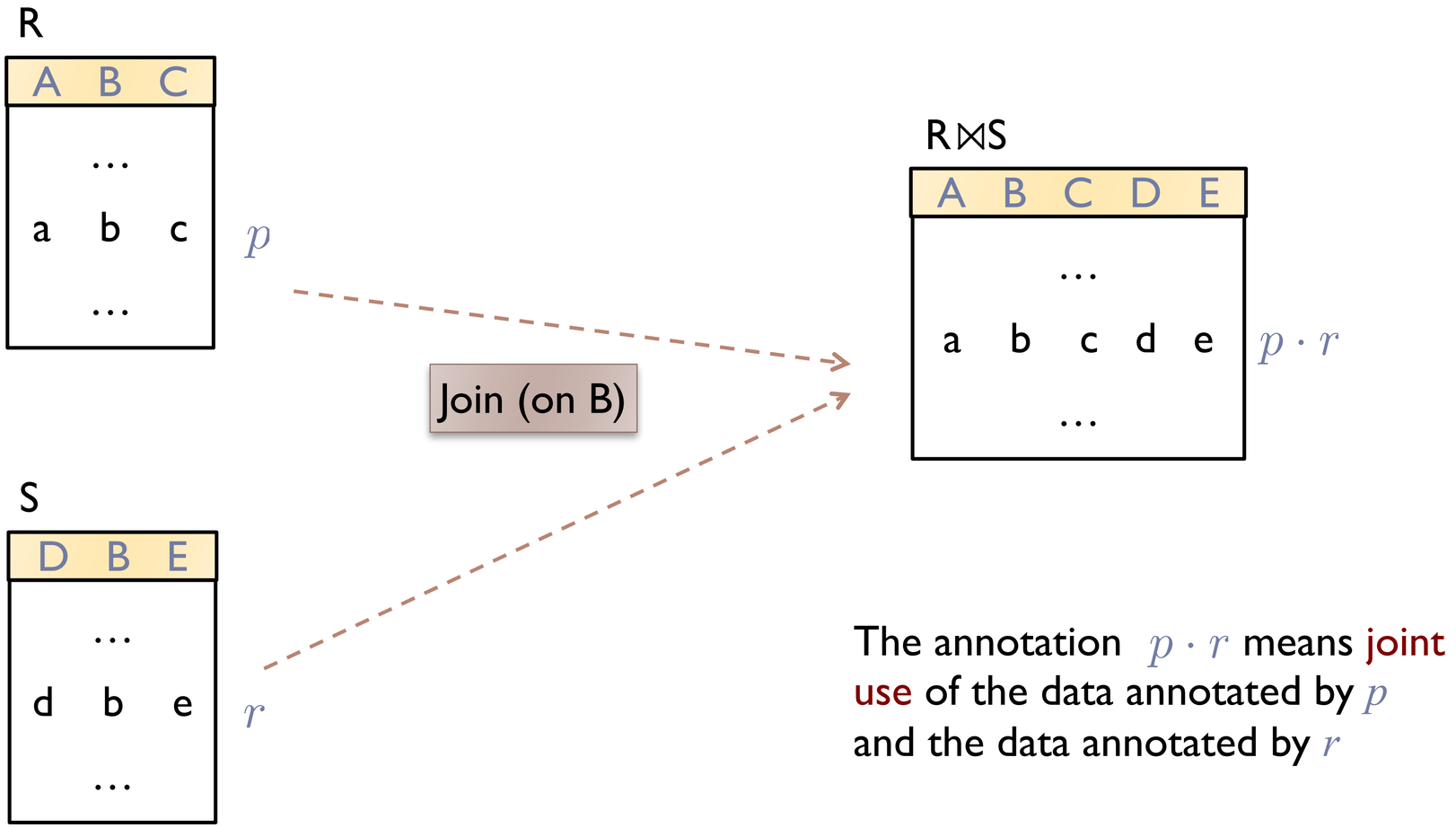
[BunemanKhannaTan. PODS 2002]

How provenance

- Identifies “witness tuples” ***and*** the operations performed on them to produce each result tuple

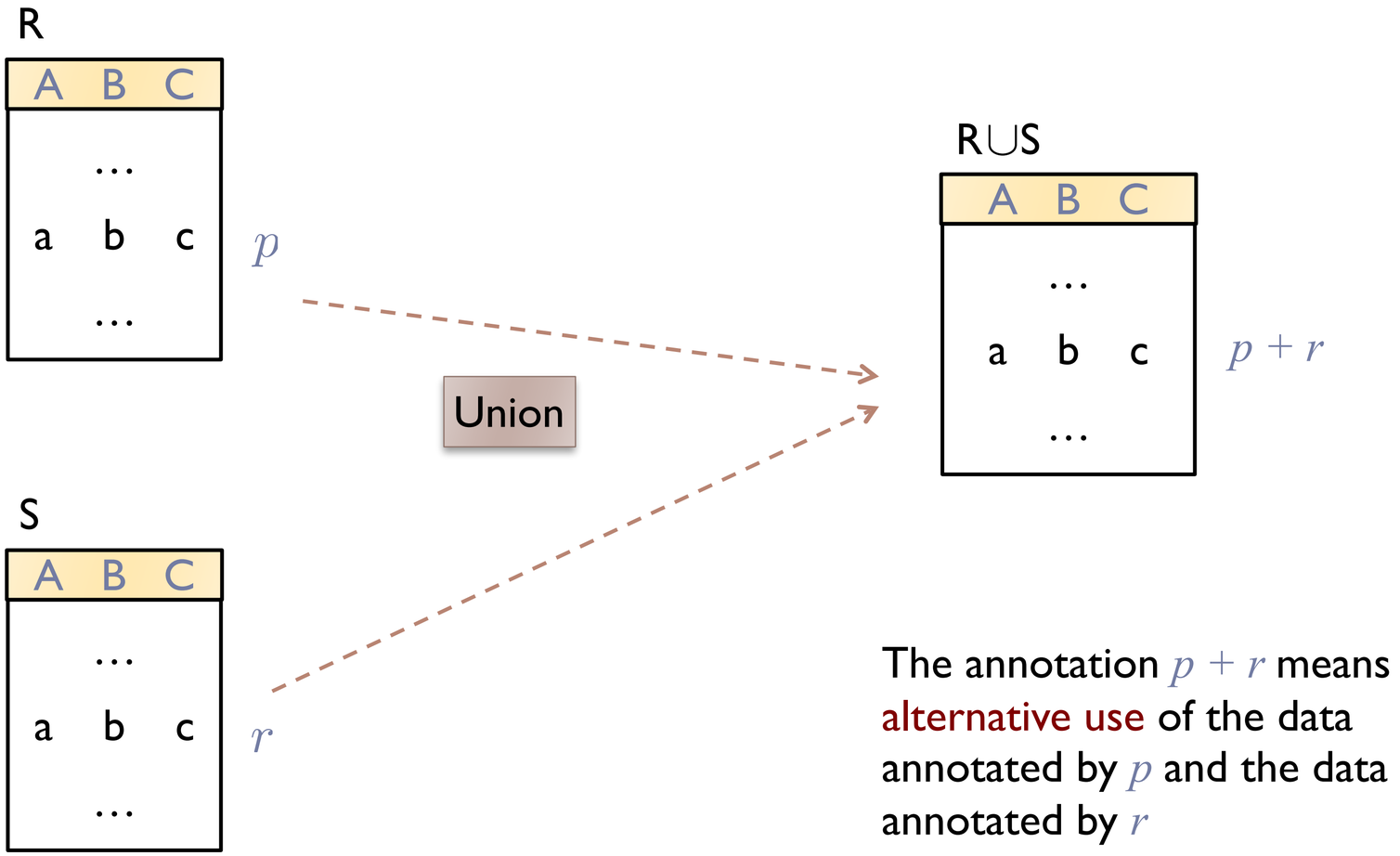
- Expresses operations using provenance semirings
  - MERGE (+): union or projection
  - JOIN (•): joins

Propagating annotations (1)



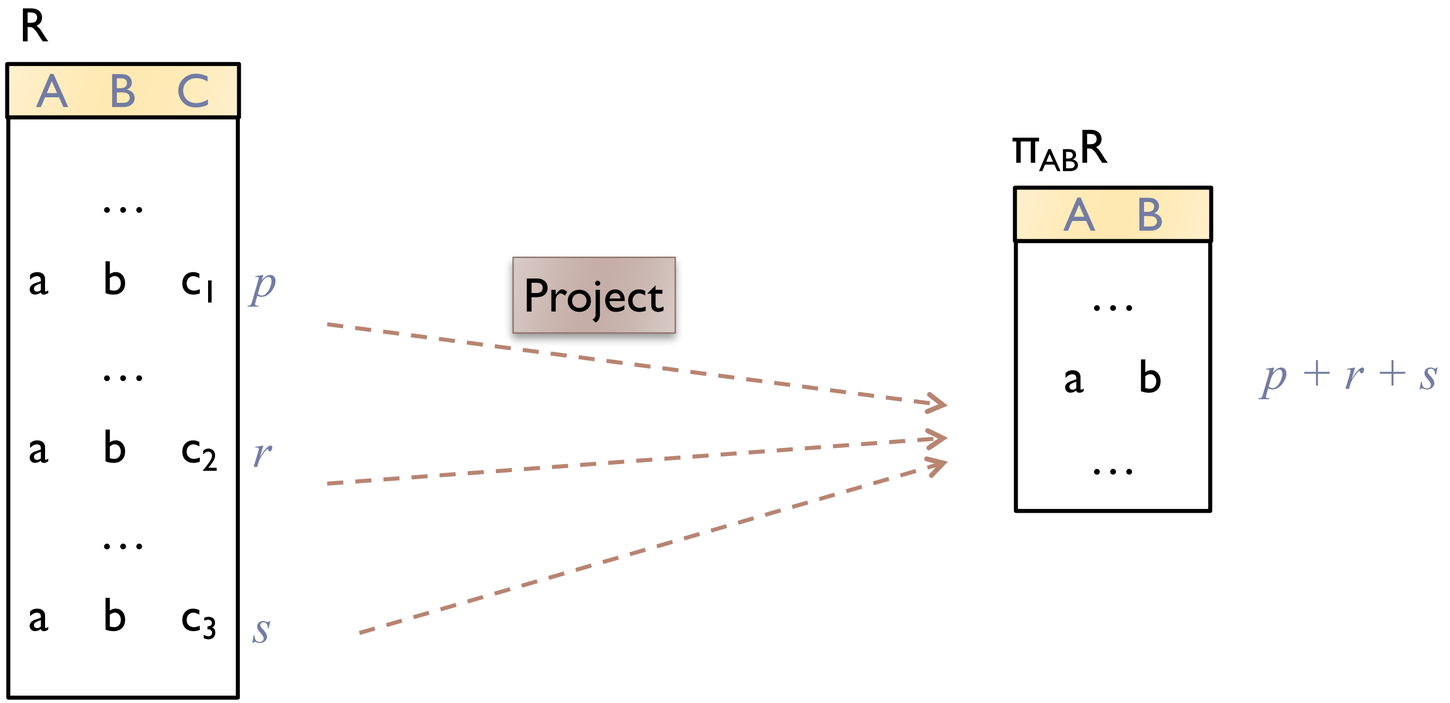
The annotation $p \cdot r$ means joint use of the data annotated by p and the data annotated by r

Propagating annotations (2)



The annotation $p + r$ means alternative use of the data annotated by p and the data annotated by r

Propagating annotations (3)



+ denotes alternative use of data

An example (SPJU)

R

A	B	C	
a	b	c	p
d	b	e	r
f	g	e	s

$$Q = \sigma_{C=e}\pi_{AC}(\pi_{AB}R \bowtie \pi_{BC}R \cup \pi_{AC}R \bowtie \pi_{BC}R)$$

A	C	
a	c	$(p \cdot p + p \cdot p) \cdot 0$
a	e	$p \cdot r \cdot 1$
d	c	$r \cdot p \cdot 0$
d	e	$(r \cdot r + r \cdot s + r \cdot r) \cdot 1$
f	e	$(s \cdot s + s \cdot r + s \cdot s) \cdot 1$

For selection, multiply with annotation 0 and 1.

Example

Agencies

	name	based_in	phone
t_1:	BayTours	San Francisco	415-1200
t_2:	HarborCruz	Santa Cruz	831-3000

ExternalTours

	name	destination	type	price
t_3:	BayTours	San Francisco	cable car	$50
t_4:	BayTours	Santa Cruz	bus	$100
t_5:	BayTours	Santa Cruz	boat	$250
t_6:	BayTours	Monterey	boat	$400
t_7:	HarborCruz	Monterey	boat	$200
t_8:	HarborCruz	Carmel	train	$90

Q:

```
SELECT destination, a.phone
FROM   Agencies a,
       (SELECT name,
               based_in AS destination
       FROM    Agencies a
       UNION
       SELECT  name, destination
       FROM    ExternalTours) e
WHERE  a.name = e.name
```

Example

Agencies

	name	based_in	phone
t_1:	BayTours	San Francisco	415-1200
t_2:	HarborCruz	Santa Cruz	831-3000

ExternalTours

	name	destination	type	price
t_3:	BayTours	San Francisco	cable car	$50
t_4:	BayTours	Santa Cruz	bus	$100
t_5:	BayTours	Santa Cruz	boat	$250
t_6:	BayTours	Monterey	boat	$400
t_7:	HarborCruz	Monterey	boat	$200
t_8:	HarborCruz	Carmel	train	$90

e

name	destination

Q:

```
SELECT destination, a.phone
FROM   Agencies a,
       (SELECT name,
               based_in AS destination
        FROM   Agencies a
        UNION
        SELECT name, destination
        FROM   ExternalTours) e
WHERE  a.name = e.name
```

Example

Agencies

	name	based_in	phone
t_1:	BayTours	San Francisco	415-1200
t_2:	HarborCruz	Santa Cruz	831-3000

ExternalTours

	name	destination	type	price
t_3:	BayTours	San Francisco	cable car	$50
t_4:	BayTours	Santa Cruz	bus	$100
t_5:	BayTours	Santa Cruz	boat	$250
t_6:	BayTours	Monterey	boat	$400
t_7:	HarborCruz	Monterey	boat	$200
t_8:	HarborCruz	Carmel	train	$90

e

	name	destination
$t_1 + t_3$	BayTours	San Francisco
$t_4 + t_5$	BayTours	Santa Cruz
t_6	BayTours	Monterey
t_7	HarborCruz	Monterey
t_8	HarborCruz	Carmel
t_2	HarborCruz	Santa Cruz

Q:

```
SELECT destination, a.phone
FROM   Agencies a,
       (SELECT name,
               based_in AS destination
        FROM   Agencies a
        UNION
        SELECT name, destination
        FROM   ExternalTours) e
WHERE  a.name = e.name
```

Example

Agencies

	name	based_in	phone
t_1:	BayTours	San Francisco	415-1200
t_2:	HarborCruz	Santa Cruz	831-3000

ExternalTours

	name	destination	type	price
t_3:	BayTours	San Francisco	cable car	$50
t_4:	BayTours	Santa Cruz	bus	$100
t_5:	BayTours	Santa Cruz	boat	$250
t_6:	BayTours	Monterey	boat	$400
t_7:	HarborCruz	Monterey	boat	$200
t_8:	HarborCruz	Carmel	train	$90

e

	name	destination
$t_1 + t_3$	BayTours	San Francisco
$t_4 + t_5$	BayTours	Santa Cruz
t_6	BayTours	Monterey
t_7	HarborCruz	Monterey
t_8	HarborCruz	Carmel
t_2	HarborCruz	Santa Cruz

Q:

```
SELECT destination, a.phone
FROM   Agencies a,
       [                                ]
WHERE  a.name = e.name
```

Example

Agencies

	name	based_in	phone
t_1:	BayTours	San Francisco	415-1200
t_2:	HarborCruz	Santa Cruz	831-3000

e

	name	destination
$t_1 + t_3$	BayTours	San Francisco
$t_4 + t_5$	BayTours	Santa Cruz
t_6	BayTours	Monterey
t_7	HarborCruz	Monterey
t_8	HarborCruz	Carmel
t_2	HarborCruz	Santa Cruz

Q:

```
SELECT destination, a.phone
FROM   Agencies a,
       
WHERE  a.name = e.name
```

RESULT

destination	phone

Example

Agencies

	name	based_in	phone
t_1:	BayTours	San Francisco	415-1200
t_2:	HarborCruz	Santa Cruz	831-3000

e

	name	destination
$t_1 + t_3$	BayTours	San Francisco
$t_4 + t_5$	BayTours	Santa Cruz
t_6	BayTours	Monterey
t_7	HarborCruz	Monterey
t_8	HarborCruz	Carmel
t_2	HarborCruz	Santa Cruz

Q:

```
SELECT destination, a.phone
FROM   Agencies a,
       
WHERE  a.name = e.name
```

RESULT

destination	phone
San Francisco	415-1200

Example

Agencies

	name	based_in	phone
t_1:	BayTours	San Francisco	415-1200
t_2:	HarborCruz	Santa Cruz	831-3000

e

	name	destination
$t_1 + t_3$	BayTours	San Francisco
$t_4 + t_5$	BayTours	Santa Cruz
t_6	BayTours	Monterey
t_7	HarborCruz	Monterey
t_8	HarborCruz	Carmel
t_2	HarborCruz	Santa Cruz

Q:

```
SELECT destination, a.phone
FROM   Agencies a,
       [                                ]
WHERE  a.name = e.name
```

RESULT

destination	phone
San Francisco	415-1200

$t_1 \cdot (t_1 + t_3)$

Example

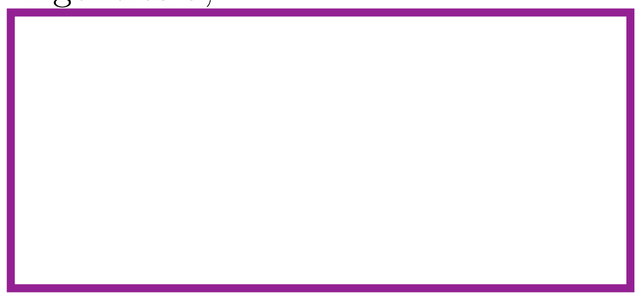
Agencies

	name	based_in	phone
t_1:	BayTours	San Francisco	415-1200
t_2:	HarborCruz	Santa Cruz	831-3000

e

	name	destination
$t_1 + t_3$	BayTours	San Francisco
$t_4 + t_5$	BayTours	Santa Cruz
t_6	BayTours	Monterey
t_7	HarborCruz	Monterey
t_8	HarborCruz	Carmel
t_2	HarborCruz	Santa Cruz

Q:

```
SELECT destination, a.phone
FROM   Agencies a,
       
WHERE  a.name = e.name
```

RESULT

destination	phone
San Francisco	415-1200
Santa Cruz	415-1200

$t_1 \cdot (t_1 + t_3)$

Example

Agencies

	name	based_in	phone
t_1:	BayTours	San Francisco	415-1200
t_2:	HarborCruz	Santa Cruz	831-3000

e

	name	destination
$t_1 + t_3$	BayTours	San Francisco
$t_4 + t_5$	BayTours	Santa Cruz
t_6	BayTours	Monterey
t_7	HarborCruz	Monterey
t_8	HarborCruz	Carmel
t_2	HarborCruz	Santa Cruz

Q:

```
SELECT destination, a.phone
FROM   Agencies a,
       [                    ]
WHERE  a.name = e.name
```

RESULT

destination	phone	
San Francisco	415-1200	$t_1 \cdot (t_1 + t_3)$
Santa Cruz	415-1200	$t_1 \cdot (t_4 + t_5)$

Example

Agencies

	name	based_in	phone
t_1:	BayTours	San Francisco	415-1200
t_2:	HarborCruz	Santa Cruz	831-3000

e

	name	destination
$t_1 + t_3$	BayTours	San Francisco
$t_4 + t_5$	BayTours	Santa Cruz
t_6	BayTours	Monterey
t_7	HarborCruz	Monterey
t_8	HarborCruz	Carmel
t_2	HarborCruz	Santa Cruz

Q:

```
SELECT destination, a.phone
FROM   Agencies a,
       e
WHERE  a.name = e.name
```

RESULT

destination	phone	
San Francisco	415-1200	$t_1 \cdot (t_1 + t_3)$
Santa Cruz	415-1200	$t_1 \cdot (t_4 + t_5)$
Monterey	415-1200	$t_1 \cdot t_6$
Monterey	831-3000	$t_2 \cdot t_7$
Carmel	831-3000	$t_2 \cdot t_8$
Santa Cruz	831-3000	t_2^2

Back to example

R

A	B	C	
a	b	c	p
d	b	e	r
f	g	e	s

Q

A	C	
a	c	$(p \cdot p + p \cdot p) \cdot 0$
a	e	$p \cdot r \cdot 1$
d	c	$r \cdot p \cdot 0$
d	e	$(r \cdot r + r \cdot s + r \cdot r) \cdot 1$
f	e	$(s \cdot s + s \cdot r + s \cdot s) \cdot 1$

Applying the laws: polynomials

R

A	B	C	
a	b	c	p
d	b	e	r
f	g	e	s

Q

A	C	
a	e	pr
d	e	$2r^2 + rs$
f	e	$rs + 2s^2$

Polynomials with coefficients in $\mathbb{N}$ and annotation tokens as indeterminates p, r, s capture a very general form of provenance

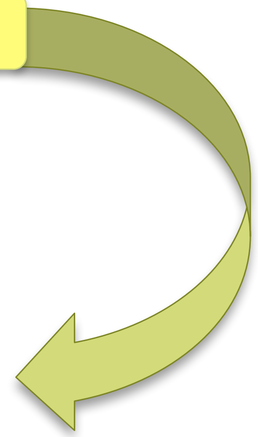
How to read this provenance

R

A	B	C	
a	b	c	p
d	b	e	r
f	g	e	s

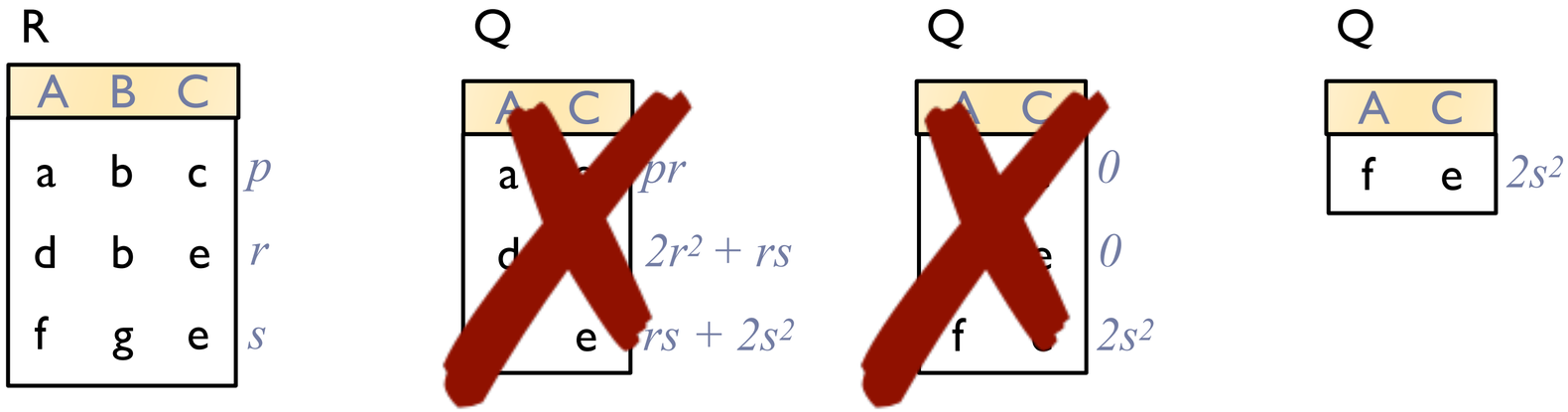
Q

A	C	
a	e	pr
d	e	$2r^2 + rs$
f	e	$rs + 2s^2$



- 3 ways to derive (d e)
- 2 of the ways use only r, but they use it twice
- the 3rd uses r once and s once

Deletion Propagation



Delete (d b e) from R

Set r to 0!

Provenance Semirings

- Space of annotations K
- **K-relations: every tuple annotated with elements from K**
- Binary operations on K
  - $\cdot$: **joint use (join)**
  - $+$: alternative use (union/projection)
- Special annotations 0 and 1 in K
  - Absent tuples ← 0
  - 1 **is a neutral annotation**

- What are the laws of $(K, +, \cdot, 0, 1)$?

Commutative Semirings?

An algebraic structure $(K, +, \bullet, 0, 1)$ where:

- K is the domain
- $+$ is associative, commutative, with 0 identity
- $\bullet$ is associative with 1 identity
- $\bullet$ distributes over $+$
- $a \bullet 0 = 0 \bullet a = 0$

semiring

- $\bullet$ is also commutative

Some useful commutative semirings

$(\mathbb{B}, \vee, \wedge, \texttt{false}, \texttt{true})$ Set Semantics

$(\mathbb{N}, +, \cdot, 0, 1)$ Bag Semantics

$(P(\Omega), \cup, \cap, \emptyset, \Omega)$ Probabilistic events

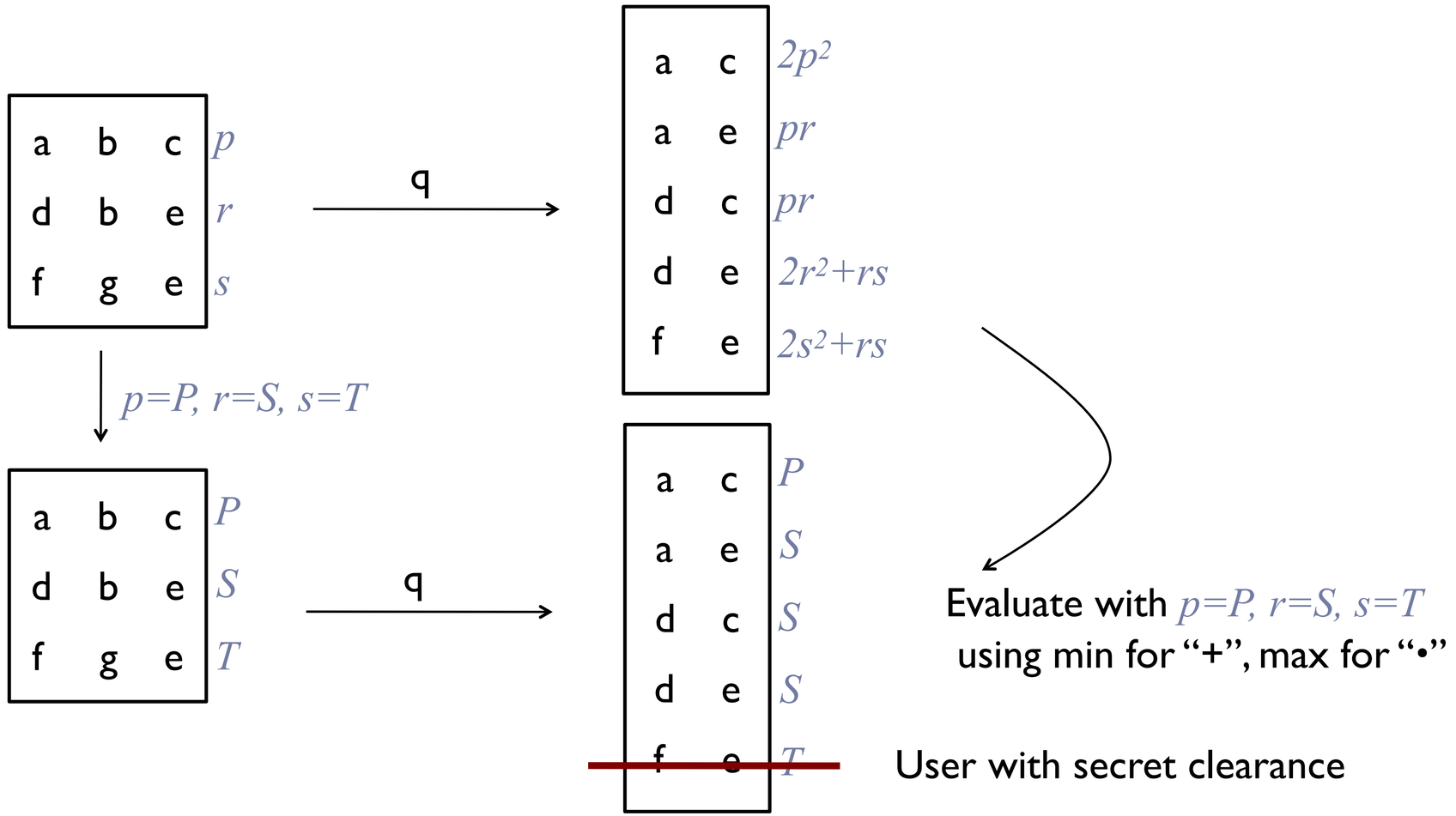
$(\mathbb{A}, \min, \max, 0, P)$
$\mathbb{A} = P < C < S < T < 0$ Access Control

Public

Top Secret

Example: access control

$(\mathbb{A}, \min, \max, 0, P)$ where $\mathbb{A} = P < C < S < T < 0$

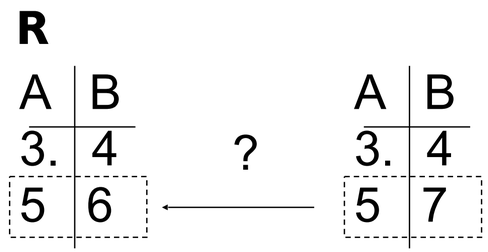


Where provenance

- Identifies “witness cells”
- Important for annotations

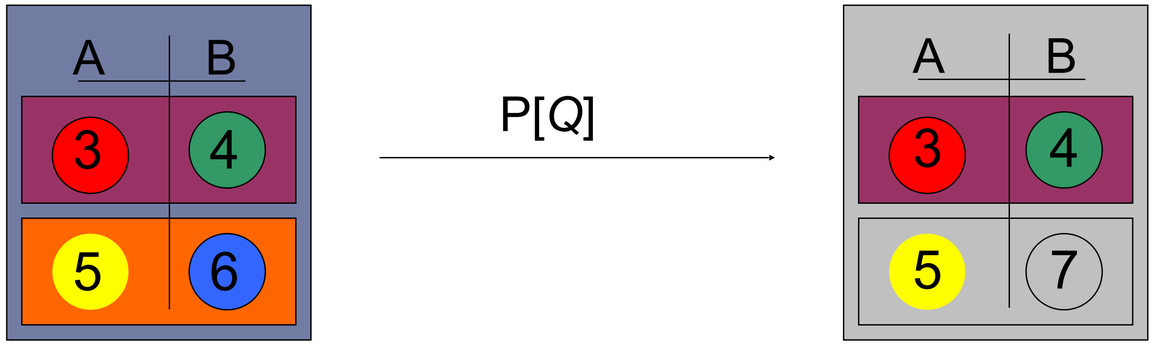
```
SELECT * FROM R WHERE A <> 5
UNION
SELECT A, 7 AS B FROM R WHERE A= 5
```

```
UPDATE R SET B=7 WHERE A=5
```



Color algebra

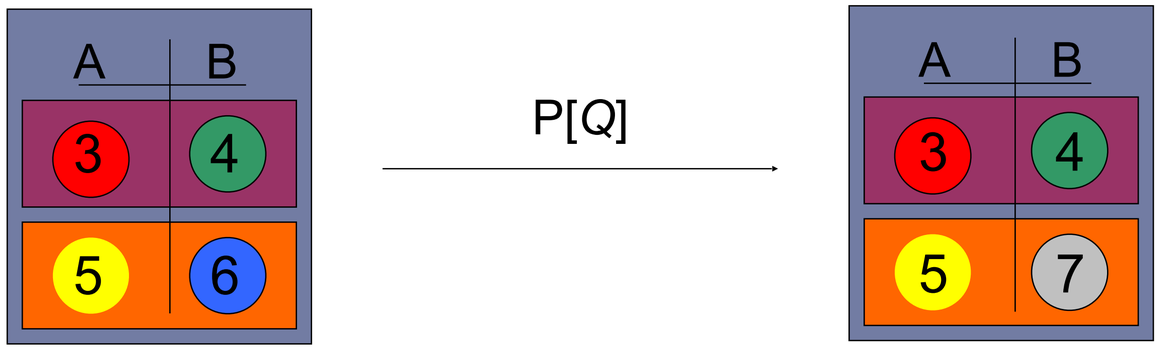
〔Geerts, Kementsietsidis, Milano 06〕



Q =

```
SELECT * FROM R WHERE A <> 5
UNION
SELECT A, 7 AS B FROM R WHERE A= 5
```

Color algebra



Q = UPDATE R SET B=7 WHERE A=5

Where provenance and semirings

R^u

A^x	B^y	C^1	
…			
a^1	b^1	c^1	p
…			

S^v

B^1	C^1	
…		
b^z	c^1	m
…		

$\pi_{AC}(\pi_{AB}R \bowtie(\pi_{BC}R \cup S))$

A^1	C^1	
…		
a^1	c^1	$u^2p^2xy^2 + uvpmxyz$
…		

1 is a neutral annotation, used when we don't bother to track data

Different annotations → Different tuples

R

A	B	C	
a	b	c	p
d	b	e^z	r
f	g	e^w	s

$\pi_C \sigma_{C=e} \pi_{AC}(\pi_{AB}R \bowtie \pi_{BC}R)$

C	
e^z	$pr+r^2$
e^w	s^2

Wrap up: issues and directions

- Archiving
- Compression
- Generalizations
  - Program Slicing [Cheney07]
- "Negative" Provenance
  - Why Not? [SIGMOD09], Artemis [PVLDB09]
- Causality